DOCUMENT RESUME

ED 246 093 TM 840 366

AUTHOR Lee, Jo Ann; And Others
TITLE The Effects of Mode of Test Administration on Test Performance.
PUB DATE Apr 84
NOTE 21p.; Paper presented at the Annual Meeting of the Eastern Psychological Association (55th, Baltimore, MD, April 12, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Adults; Comparative Analysis; *Computer Assisted Testing; *Difficulty Level; Test Anxiety; *Test Format; Testing; *Testing Problems; Test Items
IDENTIFIERS Marine Corps; *Paper and Pencil Tests

ABSTRACT

The difficulty of test items administered by paper and pencil were compared with the difficulty of the same items administered by computer. The study was conducted to determine if an interaction exists between mode of test administration and ability. An arithmetic reasoning test was constructed for this study. All examinees had taken the Armed Services Vocational Aptitude Battery (ASVAB) a short time prior to the experiment. The subject's number-correct score for the Arithmetic Reasoning subtest of the ASVAB was used as an independent estimate of ability. Regression analysis was used to test for a significant interaction between mode of administration and ability. Results indicated the computerized test to be more difficult. The anxiety level may have been higher in the computer mode, adversely affecting performance. More research is needed to corroborate the existence of significant differences between the modes. Further research is needed to identify specific factors affecting test performance in the two modes. (DWH)

The Effects of Mode of Test Administration on Test Performance

Jo Ann Lee

The University of North Carolina at Charlotte

Kathleen E. Moreno and J. B. Sympson

Navy Personnel Research and Development Center

Running Head: Mode of Test Administration

Paper presented at the Fifty-fifth Annual Meeting of the Eastern Psychological Association, Inc., Baltimore, Maryland, April 12, 1984

Abstract

This study compared the difficulty of test items administered by paper-and-pencil with the difficulty of the same items administered by computer and determined if a mode by ability interaction exists. A significant main effect for mode of administration was found. No significant mode by ability interaction was found.

The Effects of Mode of Test Administration on Test Performance

Computerized Adaptive Testing (CAT) requires a given examinee to interact individually with a computer. Test items are presented singly based upon the examinee's responses to previous items. The computer program re-estimates the examinee's ability level after he or she responds to an item then selects the next item which is most appropriate to that examinee's re-estimated ability level. The computer administers and scores the test and records the score. CAT has been made possible from the relatively recent advances in computer technology and theoretical developments in item response theory (IRT; Hambleton & Cook, 1977; Lord, 1977; Urry, 1977). Psychometric interest in adaptive testing generates from the improved measurement that such testing strategies provide, compared to conventional testing strategies (McBride, 1979; Urry, 1977). Moreover, practical reasons (Space, 1981), e.g., cost-effectiveness (Elwood, 1972) and more efficient use of labor (see, Gedye & Miller, 1969), have been impetuses for computerizing psychological tests, regardless of whether an adaptive strategy is employed or not. Although the theoretical basis is extant and the technology is available, there are still implementation questions which must be answered (see Johnson, Godin, & Bloomquist, 1981; Johnson & Johnson, 1981). The effects, if any, of computerized testing procedures on examinees' performance are not clear. There has not been a great deal of research in this area. The studies which have been conducted have provided mixed results. Studies investigating

the reliability (Katz & Dalby, 1981; Lushene, O'Neil, & Dunn, 1974) and validity (Lushene et al., 1974) of computerized versions of personality tests have obtained coefficients comparable to the paper-and-pencil forms of the tests. Research involving the use of computer devices to administer cognitive tests have provided less consistent findings. Research with the Raven Progressive Matrices Test (Rock & Nolen, 1982; Hitti, Riffer, & Stuckles, 1971) indicates that a computerized form of the test is a viable alternative to the paper-and-pencil form. Other research (see Hansen & O'Neil, 1970; Hedl, O'Neil, & Hansen, 1973; Johnson & White, 1980; Johnson & Johnson, 1981), however, suggests that interacting with a computer to complete an intelligence test may evoke a significant amount of anxiety to affect performance.

A pattern of differences between the two modes of test administration according to the specific aptitudes tested has not been found. Some examinees have performed better on verbal tests when they were administered by computer rather than by paper-and-pencil (Serwer & Stolurow, 1970; Johnson & Mihal, 1973) while other examinees have performed poorer on verbal tests administered by computer (Johnson & Mihal, 1973; Wildgrube, 1982) rather than paper-and-pencil. Still other examinees have shown no difference in performance between the two modes on verbal tests (Sachar & Fletcher, 1977) or tests which require memory retrieval (English, Reckase, & Patience, 1977; Hoffman & Lundberg, 1976). Similarly, no pattern has been found for quantitative

ability. Johnson and Mihal's (1973) subjects performed better better on quantitative tests when the tests were computer administered. In contrast, Wildgrube (1982) found no significant differences in performance between the modes for arithmetic reasoning. Studies involving other nonverbal tests, e.g., figural reasoning (Wildgrube, 1982) and analytical processing (Sachar & Fletcher, 1977) have also produced mixed results.

In summary, the effects of mode of test presentation on performance are not clear. Conflicting findings in previous research might be due to differences in methodology. Finding differences between modes might depend upon test content (e.g., personality tests vs. cognitive tests or easy tests vs. difficult tests or verbal test vs. quantitative tests), the population tested (e.g., blacks vs. whites or naive subjects vs. experienced subjects), or the design of the study (e.g., repeated measures vs. independent groups or sample size).

The purpose of this study was (1) to compare the mean difficulty of test items which were administered by paper-and-pencil with the mean difficulty of the same items administered by computer and (2) to determine if an interaction between mode of test administration and ability exists.

Methods

Subjects

Subjects were 654 male Marine Corps recruits between the ages of 18 and 25, stationed at the Marine Corps Recruit Depot (MCRD), San Diego, California. The paper-and-pencil test was administered to 334 recruits and the computerized test was administered to 320 recruits.

Procedure

A 30-item arithmetic reasoning test was constructed for this study. The number of items that an examinee answered correctly on the experimental arithmetic reasoning test (EXP-AR) was the dependent variable. In addition, all subjects had taken the Armed Services Vocational Aptitude Battery (ASVAB) approximately two weeks to six months prior to the experimental test. A given subject's number-correct score for the Arithmetic Reasoning subtest of the ASVAB (ASVAB-AR) was used as an independent estimate of that subject's arithmetic reasoning ability.

EXP-AR was administered to participants approximately 24 hours after their arrival at the MCRD receiving barracks. Each subject was randomly assigned to one of the two modes of test administration.

Subjects in the paper-and-pencil mode were tested in groups of 4 to 10. Each subject was given a test booklet containing test instructions, three sample questions, and the 30 test items. There were approximately eight items per page. Item responses were recorded on an answer sheet. It was possible for examinees to refer to previous items and to change their answers.

Subjects in the computer mode were tested in groups of four, using cathode-ray tube terminals. Test instructions were presented by the computer. The instructions were written to be as similar as possible to those given in the paper-and-pencil mode, except additional instructions on the use of the computer terminal were given. The same three sample questions that were given in the paper-and-pencil

mode were administered by the computer. Each sample and test item was displayed individually on the screen. Keys used to enter item responses were specially labelled with bold, black letters on a white background. It was not possible for examinees in the computer mode to refer to previous items nor to change their answers once the answer had been entered on the keyboard and recorded by the computer.

Time limits were not imposed and omitting of items was not allowed in either mode of administration.

Results

Sixty-nine subjects were deleted from the original sample because of incomplete data. The final sample size was 585, with 300 in the paper-and-pencil mode and 285 in the computer mode.

Linear regression analysis was used to perform an analysis of covariance, with EXP-AR as the dependent variable, mode of test administration as the independent variable, and ASVAB-AR as the covariate.

A significant main effect for mode of administration was found ($p<.01$).

As shown in Table 1, the mean ASVAB-AR number-correct scores for the two groups were very close in value. This indicates that, on the basis of arithmetic reasoning ability, random assignment to groups was successful. Mean number-correct scores for the experimental test given under the two modes of administration were significantly different from each other. Regression analysis was used to further investigate this difference.

The following regression model was used to test for a significant interaction between mode of administration and ability:

$$E(Y) = B_0 + B_1X + B_2M + B_3(MX),$$

where Y was the EXP-AR score, X was the pre-enlistment ASVAB-AR score (the covariate), M was +1 if the examinee was in the paper-and-pencil group and -1 if the examinee was in the computer group, and MX was the product of M and X (the interaction term). The B symbolizes raw-score regression weights.

Results showed B_3 was not significantly different than zero, indicating that there was no significant interaction between ability and mode-of-administration. The effect of ability (as measured by ASVAB-AR) on the dependent measure (EXP-AR) was the same regardless of mode of test administration. Therefore, the following model was the appropriate one to fit:

$$E(Y) = B_0 + B_1X + B_2M.$$

The multiple regression coefficient for this model was .75, with B_1=.8102, F=747.702, and B_2=.5133, F=10.793, $p<.01$. Since B_2 was significantly different than zero, this indicates the presence of a main effect for mode of test administration.

Information from the regression analysis was used to obtain the two within group regressions of EXP-AR on ASVAB-AR. Figure 1 shows a plot of these regression lines, superimposed upon a scatterplot in which ASVAB-AR number-correct score is on the horizontal axis and EXP-AR is on the vertical axis. The difference between the intercepts for these two parallel lines was 1.0277. The means for the paper-

and-pencil group and the computer group, "adjusted for" the ASVAB-AR covariate, were 19.31 and 18.28, respectively. Therefore, subjects in the paper-and-pencil mode of test administration scored, on the average, 1.03 raw-score points above the subjects in the computer mode.

Item analysis was performed to determine if the effect of mode of test administration was the same over all test items or if some items were affected more than others. Figure 2 shows a scatterplot of the p-values for the two groups. Twenty-one of the 30 items were more difficult in the computer mode, while only three were more difficult in the paper-and-pencil mode. The remaining six items were of approximately equivalent difficulty. This result shows that item difficulty was affected by mode of administration and that this effect was fairly constant across items.

Implications and Conclusions

The obtained main effect by mode was unexpected. It is not obvious what caused the computerized test to be more difficult. The anxiety level may have been significantly higher in the computer mode, which adversely affected performance (see Hansen & O'Neil, 1970; Hedl, O'Neil, & Hansen, 1973; Johnson & Johnson, 1981). More training in the use of computers to alleviate possible computer-evoked anxiety is suggested in future research and applications.

On the other hand, past research has failed to consistently find significant differences between the two modes of presentation, without specifically controlling for anxiety. Moreover, the pattern

of differences between the modes across different abilities has not been consistently replicated across studies. Alternatively, the number of items present at a given time (e.g., eight in the paper-and-pencil mode vs. one in the computerized mode) may significantly affect performance on certain types of items (see Hoffman & Lundberg, 1976). The results from the current study indicate that more research is needed to corroborate the existence of significant differences between the modes. Further research is especially needed to identify the specific factors affecting test performance in the two modes.

References

Elwood, D. L. Test retest reliability and cost analyses of automated and face to face intelligence testing. International Journal of Man-Machine Studies, 1972, 4, 1-22.

English, R. A., Reckase, M. D., & Patience, W. M. Application of tailored testing to achievement measurement. Behavior Research Methods & Instrumentation, 1977, 9, 158-161.

Gedye, J. L., & Miller, E. The automation of psychological assessment. International Journal of Man-Machine Studies, 1969, 1, 237-262.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-94.

Hansen, D. H., & O'Neil, H. F. Empirical investigations versus anecdotal observations concerning anxiety and computer assisted instruction. Journal of School Psychology, 1970, 8, 315-316.

Hedl, J. J., O'Neil, H. F., & Hansen, D. H. Affective reactions toward computer-based intelligence testing. Journal of Consulting and Clinical Psychology, 1973, 40, 217-222.

Hitti, F. J., Riffer, R. L., & Stuckless, E. R. Computer-managed testing: A feasibility study with deaf students. National Technical Institute for the Deaf, July 1971.

Hoffman, K. I., & Lundberg, G. D. A comparison of computer-monitored group tests with paper-and-pencil tests. Educational and Psychological Measurement, 1976, 36, 791-809.

Johnson, D. F., & Mihal, W. L. Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 1973. 28, 694-699.

Johnson, D. F., & White, C. B. Effects of training on computerized test performance in the elderly. Journal of Applied Psychology, 1980, 65, 357-358.

Johnson, J. H., Godin, S. W., & Bloomquist, M. L. Human factors engineering in computerized mental health care delivery. Behavior Research Methods & Instrumentation, 1981, 13, 425-429.

Johnson, J. H., & Johnson, K. N. Psychological considerations related to the development of computerized testing stations. Behavior Research Methods & Instrumentation, 1981, 13, 421-424.

Katz, L., & Dalby, J. T. Computer-assisted and traditional psychological assessment of elementary-school-age children. Contemporary Educational Psychology, 1981, 6, 314-322.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-137.

Lushene, R. E., O'Neil, H. F., Jr., & Dunn, T. Equivalent validity of a completely computerized MMPI. Journal of Personality Assessment, 1974, 34, 353-361.

McBride, J. R. Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Rock, D. L., & Nolen, P. A. Comparison of the standard and computerized versions of the Raven Coloured Progressive Matrices Test. Perceptual and Motor Skills, 1982, 54, 40-42.

Sachar, J. D., & Fletcher, J. D. Administering paper-and-pencil tests by computer, or the medium is not always the message. Proceedings

Serwer, B. L., & Stolurow, L. M. Computer-assisted learning in language arts. Elementary English, 1970, 47, 641-650.

Space, L. G. The computer as psychometrician. Behavior Research Methods & Instrumentation, 1981, 13, 595-606.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Education Measurement, 1977, 145, 181-195.

Wildgrube, W. Computerized testing in the German Federal Armed Forces (FAF)--Empirical approaches. Presented at the 1982 Computerized Adaptive Testing Conference, Spring Hill, Minnesota, July 1982.

Table 1

Descriptive Statistics for Experimental AR and ASVAB AR Broken Down by Experimental Group

Variable	N of Cases	Mean	Std. Dev.	T Value	2-Tail Prob.
Experimental AR					
Paper-and-Pencil Group	300	19.31	5.62	2.19	.03
Computer Group	285	18.27	5.81		
ASVAB AR					
Paper-and-Pencil Group	300	20.66	5.27	.02	.98
Computer Group	285	20.65	5.31		

Figure 1. Performance (number-correct scores) on the Experimental-Arithmetic Reasoning Test (Experimental AR) as a function of performance (number-correct scores) on the Armed Services Vocational Aptitude Battery-Arithmetic Reasoning subtest (ASVAB AR). The solid line (——) is the regression line for the paper-and-pencil mode of test administration; the dashed line (---) is the regression line for the computer mode of test administration. The circles (o) represent data points for the paper-and-pencil mode; the crosses (x) represent data points for the computer mode.

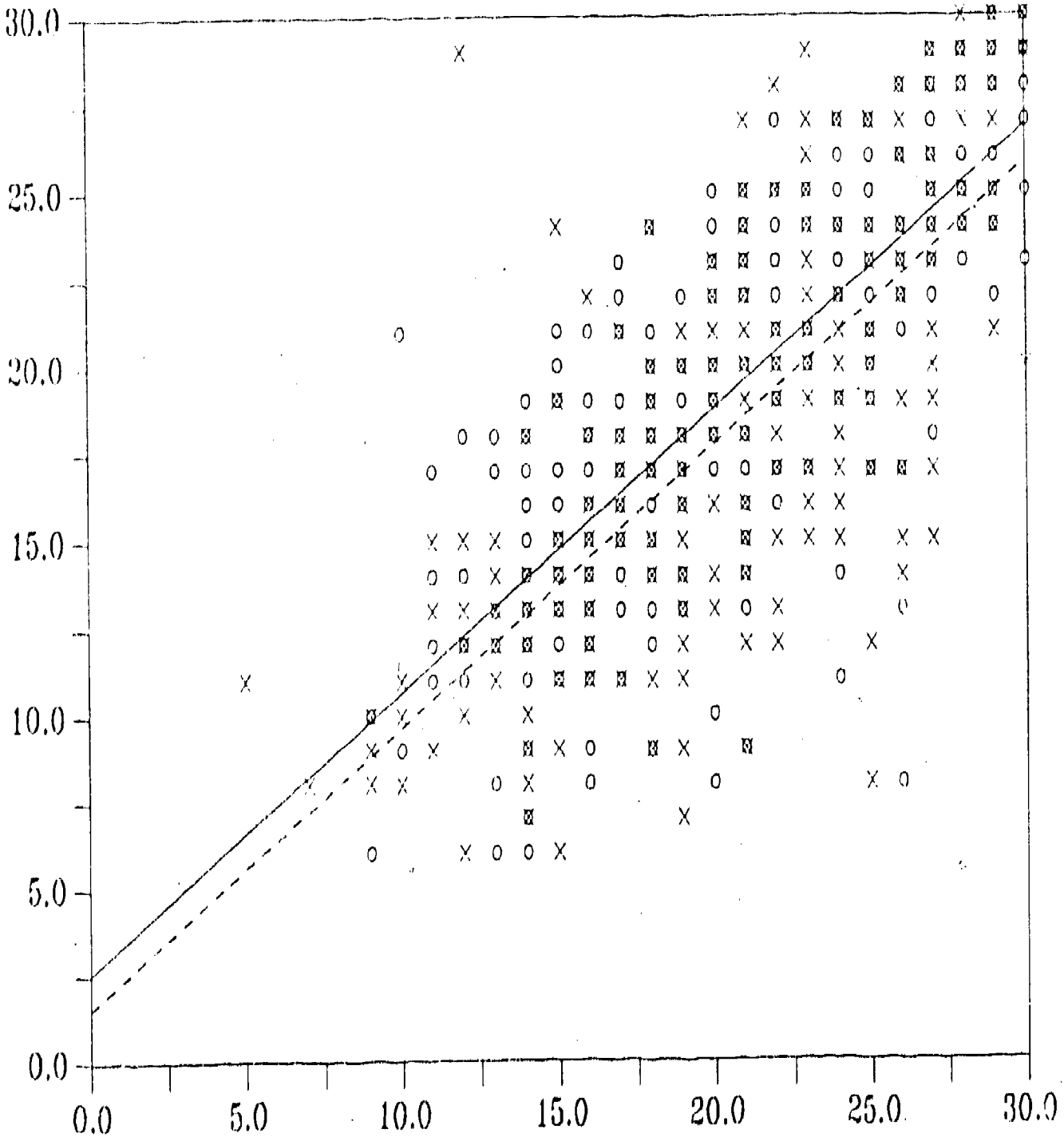
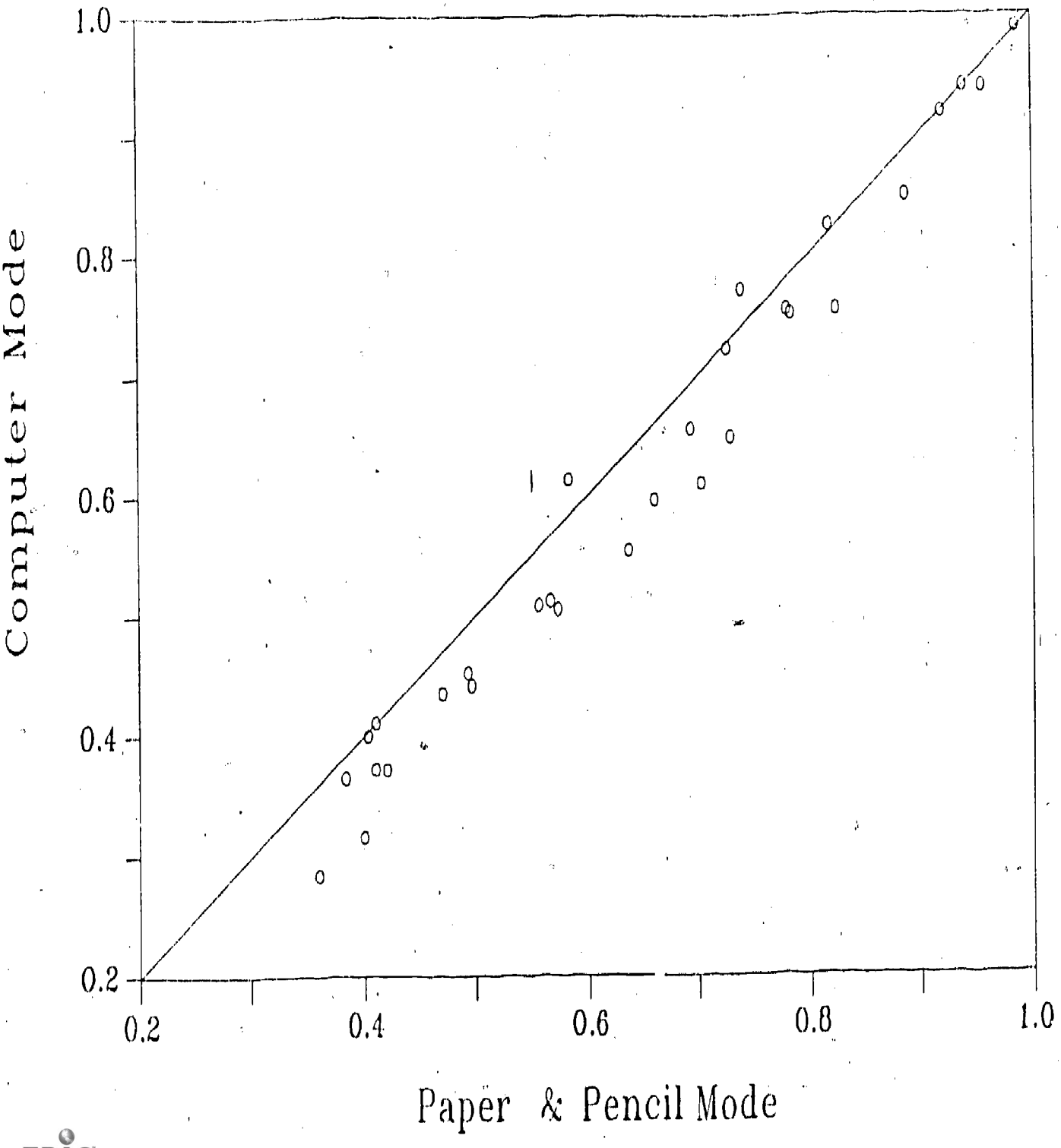

Experimental AR
30.0
25.0
20.0
15.0
10.0
5.0
0.0
0.0
5.0
10.0
15.0
20.0
25.0
30.0
ASVAB AR

Figure 2. A scatterplot of the item difficulty indices (p-values) for the paper-and-pencil mode of test administration and the computer mode of test administration.


1.0
0.8
0.6
0.4
0.2
Computer Mode
0.2
0.4
0.6
0.8
1.0
Paper & Pencil Mode